ISSN 1330-3651(Print), ISSN 1848-6339 (Online)
DOI: 10.17559/TV-20150314134638

RESEARCH ON MODEL OF NETWORK INFORMATION EXTRACTION BASED ON IMPROVED TOPIC-FOCUSED WEB CRAWLER KEY TECHNOLOGY

Mo Chen, Xiao-Ping Yang

Original scientific paper

This research has caught researchers' wide attention for extracting network information exactly with the arrival of the big data era characterized by semi-structured or unstructured text. This paper proposes a model of network information extraction based on improved topic-focused web crawler key technology taking Web news as object of extraction. The authors elaborate main function, method and technology on every layer of the model in detail, which have been used or completed, and focuses on how to extract network information efficiently oriented topic from a large number of Web news instances, in order to explore a research method for network information extraction. The experimental results show the feasibility, validity and superiority of the model design and play a very important role in constructing topic-focused Web news corpus so as to provide a real-time data source for trust analysis, currency analysis, hot topic detection, topic evolution tracking of Web news.

Keywords: *network information extraction; relativity calculation; search strategy; topic-focused web crawler*

Istraživanje modela izlučivanja mrežnih informacija utemeljenog na poboljšanoj tehnologiji tematski usmjerenog pretraživača

Izvorni znanstveni članak

U ovom su istraživanju istraživači svu svoju pažnju usmjerili na skupljanje informacija s mreže upravo u vrijeme postojanja ogromne količine podataka u polustrukturiranim ili nestrukturiranim tekstovima. U radu se predlaže model skupljanja informacija s mreže temeljen na poboljšanoj tehnologiji pretraživača mreže (web crawler) usmjerenog na izabrano područje, uzimajući obavijesti na Webu kao predmet istraživanja. Autori detaljno analiziraju glavnu funkciju, metodu i tehnologiju na svakoj razini modela i usredotočuju se na način kako učinkovito s mreže iz ogromnog broja podataka pronaći potrebna saznanja o zadanoj temi kako bi istražili analiziranu metodu za dobivanje informacija s mreže. Eksperimentalni rezultati pokazuju izvedivost, valjanost i superiornost dizajna modela, igraju važnu ulogu u sastavljanju korpusa podataka pronađenih na mreži iz odabranog područja i predstavljaju izvor aktualnih informacija za pouzdanu analizu, istraživanje aktualnih tema i praćenje razvoja događanja na Webu.

Ključne riječi: *pretraživač mreže usmjeren na izabrano područje; proračun relativnosti; skupljanje podataka s mreže; strategija pretraživanja*

1 Introduction

With the arrival of the big data era, Internet and the field of information technology have developed a challenging stage so far. According to survey of TeckTarget that is a global leading professional IT network media [1], it has shown that the number of enterprises' data has broken through TB level with the development of Internet, social media, business and other fields. Based on data existed and existing, people should think how to acquire, manage and analyse complicated network data characterized by semi-structured or unstructured text, which have shown a tendency of explosive growth [2], nevertheless, in whole process of cognizing network data, extracting network information exactly and effectively is the critical and important link.

In a mass of network data, the number of the Web news released has reached PB level [3], which shows the 4V features of the big data, it is volume, variety, velocity and value [4]. Based on these features above, Web news should reflect high reliability and currency, on the basis of which the event of hot topic should also be detected quickly and its path of evolution should be tracked accurately. However, the precondition of acquiring analysis results above needs real-time data source, therefore, it has become an urgent problem solved to construct topic-focused Web news corpus so as to provide real-time data source for trust analysis, currency analysis, hot topic detection, topic evolution tracking of Web news.

This paper proposes a model of network information extraction mainly containing four layers based on improved topic-focused web crawler key technology taking Web news as object of extraction. The author elaborates main function, method and technology mainly on every layer of the model in detail, which have been used or completed, and focuses on how to extract network information efficiently focused on topic from massive Web news instances. This process of research does key contribution for exploring a method for network information extraction, these experimental results show the feasibility, validity and superiority of the extraction model design and implementation.

2 Related works

In recent several years, some scholars have conducted some research about network information extraction method using different theory and technology. For example, Wu Jiagao and others survey research on the method of network information extraction based on the character of the loose Chinese text structure and flexible grammatical peculiarity [5]. In this paper, a combination method of syntactic analysis and Hidden Markov Model for extracting network information is proposed, the experiment has shown that the method has higher precision and recall than normal algorithm. Zhang Hongli and others survey research on the method of network information extraction based on the requirement of extracting the required information from mass data efficiently and accurately for users [6]. In this paper, a method is proposed based on MapReduce for network information extraction facing the challenges posed by large-scale computing, the simulation results of experiments show that the method has high efficiency and good adaptability taking the extraction of vast Taobao's data sources as an example. Li Wen and others survey

research on the method of network information extraction based on the application of search engine to XML technology [7]. In this paper, a web information extraction model is proposed based on XML and DOM technology. The stages of data acquisition, webpage optimization, extraction rule generation and information extraction are analysed in detail, those works are related to author's research direction of network information extraction and application.

In recent several years, some scholars have also conducted certain research about technology and method of crawling web information using topic-focused web crawler. For example, Du Yajun surveys research on the strategy of crawling web information using topic-focused web crawler [8]. In this paper, a strategy of understanding, cooperating and competing is proposed based on concept context graph for topic-focused web crawler. Xie Zhijun surveys research on the method of crawling web information using topic-focused web crawler based on the requirement of collecting data resources for the topic-oriented user's query [9]. In this paper, an approach of crawling web information is proposed for topic-focused web crawler based on HMM. The results of experiments show that this method can capture a large number of high quality webpages related to target topics, and its crawling efficiency topic-focused is better than Best-First topic-focused web crawler. Bai Yuzhao surveys research on the method of crawling web information using topic-focused web crawler [10]. This paper proposes a method of crawling web information based on probability model. The experimental results show that this method can gather more topics related to webpages by retrieving less webpages and has a better average topic relativity. Those works were related to author's research direction of topic-focused web crawler key technology and application.

Based on the analysis of the related research on network information extraction method and topic-focused web crawler key technology, experts and scholars have studied in two directions, but the research of constructing a network information extraction model based on topic-focused web crawler key technology taking Web news as an object of extraction according to its information trait is less. Therefore, this paper proposes a model of network information extraction based on topic-focused web crawler key technology mainly, in order to explore how to extract network information accurately.

3 Notations and our problem definition

At present, the universal search engine has better performance in conducting common users' searching request, but facing the increasing tendency of substantive webpages and personalized searching request, it has many shortcomings in the situation of webpage content searched real-time updating and emerges the problem of lower precision and recall [11]. Based on its shortcomings and problems emerging, the topic-focused search engine oriented on specific domain emerges as the time requires, which has become one of the major development trends in search engine application direction, nevertheless, the designing of topic-focused web crawler is the core of topic-focused search engine implementation.

With the rapid development of information technology and network technology, there are many types of network information, such as short text of micro blog, short, moderate or long text of Web news, long text of document and so on, while the biggest difference is structure of text content among them. In this paper, the author selects Web news as the object of extraction in view of ensuring high adaptability that the model of network information extraction based on improved topic-focused web crawler key technology should have and further propose the improved strategy of extracting information, in order to achieve the ideal effect of network information extraction based on topic-focused web crawler in the aspects of extraction precision and so on. This research will provide scientific method for constructing and validating the model of network information extraction.

In this section, the author provides definitions used in model and algorithms based on the practical value and application direction of Web news extraction. Let *NewsSet* be a set of Web news, the model of network information extraction will extract Web news elements from this set containing Web news URLs according to search keywords. Let *UrlSet* be a set of initial Web news URL, the model of network information extraction will define topics searched and extract Web news elements from this set containing Web news URLs according to search keywords. Let *TopKeyWordSet* be a set of initial Web news topic keywords, the model of network information extraction will define topics searched by combining it and *UrlSet*. Let *SearchKeyWordSet* be a set of Web news search keywords, the model of network information extraction will extract Web news elements according to it.

Definition 3.2.1: Given a set of *NewsSet*, it can denote using $NewsSet = \{ns_1, ns_2, ns_3, \ldots, ns_{i-1}, ns_i, ns_{i+1}, \ldots, ns_n\}$, the range of i is between one and n. ns_i contains hyperlinks, which can denote using $HyperLinkSet = \{hls_{i1}, hls_{i2}, hls_{i3}, \ldots, hls_{i(j-1)}, hls_{ij}, hls_{i(j+1)}, \ldots, hls_{im}\}$, hls_{ij} represents the j hyperlink of ns_i in *HyperLinkSet*, the range of i is between one and n, the range of j is between one and m.

Definition 3.2.2: Given a set of UrlSet, it can denote using $UrlSet = \{us_1, us_2, us_3, \ldots, us_{i-1}, us_i, us_{i+1}, \ldots, us_n\}$, us_i represents the i element of Web news in *UrlSet*, the range of i is between one and n. If the element of Web news is from webpage $Page_i$, then us_i can denote using $<url_i, title_i, pubtime_i, pubsource_i, content_i>$, url_i represents the address of $Page_i$, $title_i$ represents the title of Web news, $pubtime_i$ represents the releasing time of Web news, $pubsource_i$ represents the releasing source of Web news, $content_i$ represents the text content of Web news.

Definition 3.2.3: Given a set of *TopKeyWordSet*, it can denote using $TopKeyWordSet = \{tkws_1, tkws_2, tkws_3, \ldots, tkws_{i-1}, tkws_i, tkws_{i+1}, \ldots, tkws_n\}$, $tkws_i$ represents the i topic keyword of initial Web news topic keywords in *TopKeyWordSet*, the range of i is between one and n. $tkws_i.wordvalue$ stores topic keyword, $tkws_i.weightvalue$ stores its value of weight set.

Definition 3.2.4: Given a set of *SearchKeyWordSet*, it is deduced by combining *UrlSet* and *TopKeyWordSet*, it can denote using $SearchKeyWordSet = \{skws_1, skws_2, skws_3, \ldots, skws_{i-1}, skws_i, skws_{i+1}, \ldots, skws_n\}$, $skws_i$

represents the *i* search keyword in *SearchKeyWordSet*, *i* may be bigger than the number of initial Web news topic keywords in *TopKeyWordSet*. $skws_i.wordvalue$ stores search keyword, $skws_i.weightvalue$ stores its value of weight set.

Definition 3.2.5: Given two queues of URLs, it can denote using $InitialUrlQueue = \{iuq_1, iuq_2, iuq_3, \ldots, iuq_{i-1}, iuq_i, iuq_{i+1}, \ldots, iuq_n\}$ and $WaitingUrlQueue = \{wuq_1, wuq_2, wuq_3, \ldots, wuq_{i-1}, wuq_i, wuq_{i+1}, \ldots, wuq_n\}$ respectively, iuq_i represents the i element of initial URL queue from the front of queue to the rear of queue, wuq_i represents the i element of waiting URL queue from the front of queue to the rear of queue, the range of i is between zero and n.

Definition 3.2.6: Given three sets of *NewsSet*, *UrlSet* and *TopKeyWordSet*, the problem solved by the model of network information extraction based on improved topic-focused web crawler key technology is to extract top k elements of Web news containing its every data item, the results of extraction can denote using $TopWebNews = \{twn_1, twn_2, twn_3, \ldots, twn_{i-1}, twn_i, twn_{i+1}, \ldots, tws_k\}$, which is an ordered set of top *k* Web news elements, twn_i represents the *i* element of the Web news extraction results in *TopWebNews*, the range of i is between one and *k*. $twn_i.url$ stores the url of Web news, $twn_i.title$ stores the title of Web news, $twn_i.pubtime$ stores releasing time of Web news, $twn_i.pubsource$ stores releasing source of Web news, $twn_i.content$ stores text content of Web news, $twn_i.dividedtitle$ stores words divided for $twn_i.title$, $twn_i.dividedcontent$ stores words divided for $twn_i.content$, $twn_i.contentkeyword$ stores top keywords of $twn_i.content$, $twn_i.relativity$ value stores value of relativity related to topics, $twn_i.parenturl$ stores url of parent level, $twn_i.systemtime$ stores system time of extracting Web news element.

4 The design of network information extraction model

In the era background of big data development, it has become an important research direction to extract network information exactly in Web text mining field through the process of defining extraction targets, extracting valuable network information, filtering noise information and applying information extracted and so on. Based on this process, the model of network information extraction based on topic-focused web crawler key technology taking Web news as object of extraction is divided into four layers, which include definition layer of Web news topics, extraction layer of Web news elements, filter layer of Web news elements and application layer of Web news extraction results. As shown in Fig. 1, it displays flow process and core tasks in every layer of this model.

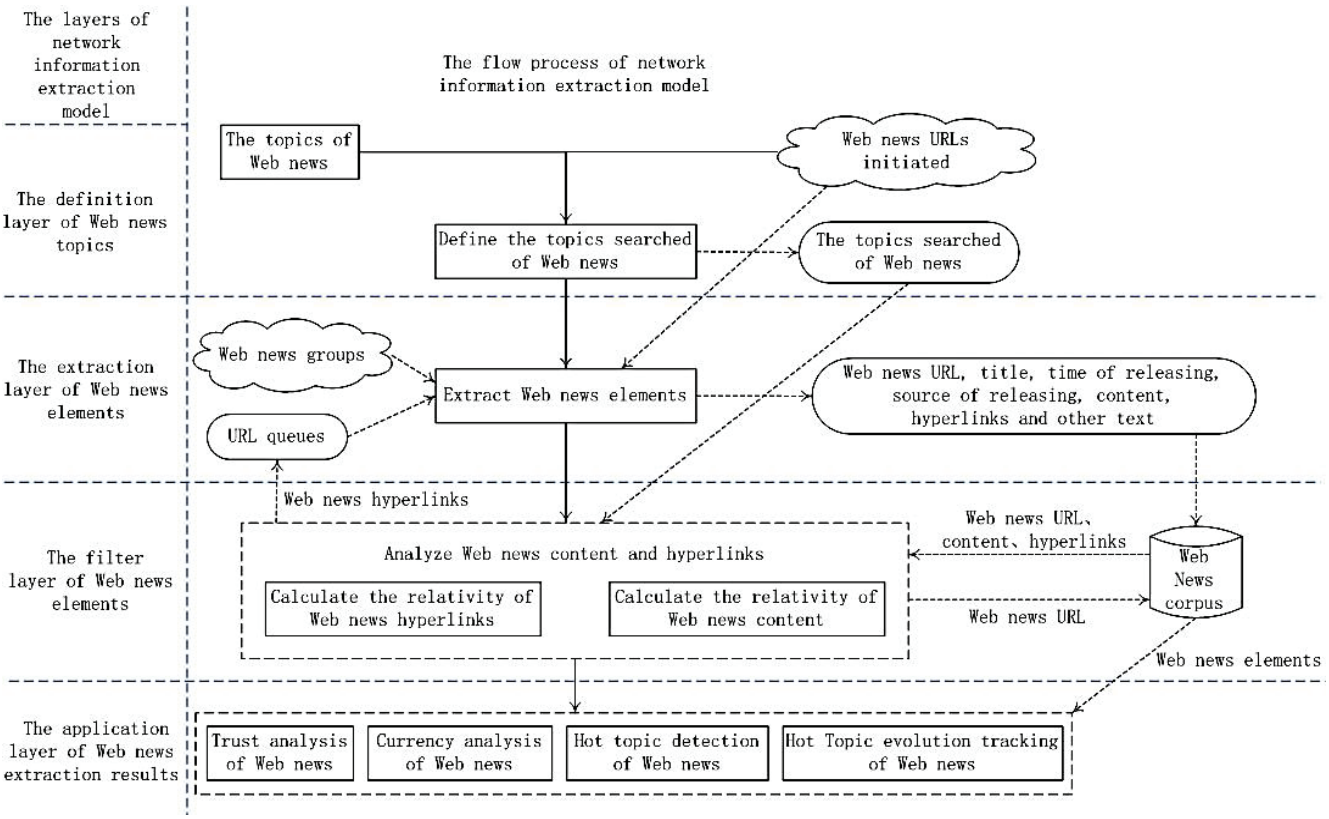

Figure 1 The model of network information extraction

4.1 The definition layer of Web news topics

The definition layer of Web news topics is mainly responsible for defining the topics searched of Web news according to *UrlSet* and *TopKeyWordSet*, which finally is and gives different value of weight for different keywords. The data source of keywords extraction is from inputting of users and Web news URL initiated, the denoted using *SearchKeyWordSet* and is used in the filter layer of Web news elements.

The precondition of topic-focused extraction is to define the topics of Web news. This paper describes crawling target using the method of keywords extraction keywords are input through consulting experts in the field of this topic and are set corresponding value of weight. According to training samples provided by initial Web

news URL, the process of keywords extraction can automatically get a set of keywords that represents topics and compute its value of weight using the algorithm based on improved TF-IDF formulae. Finally, a set of more personalized and higher precision keywords and the corresponding value of weight are acquired through training samples provided by initial Web news URL, which are also guided by keywords input from users.

The improved formula considers the importance of the same words in different categories and allocates value of weight by making a distinction among them. The value of weight *Weight*(*KeyWord*, *Document*) computation formula in document *Document* for keyword *KeyWord* is shown as follows.

$$Weight(KeyWord, Document) = \frac{F(KeyWord, Document) \times LOG\left(\frac{N}{n} + 0{,}1\right) \times Weight(KeyWord, Class)}{\sqrt{\sum_{KeyWord \in S} \left[F(KeyWord, Document) \times LOG\left(\frac{N}{n} + 0{,}1\right) \times Weight(KeyWord, Class)\right]^2}} \quad (1)$$

As shown in Eq. (1), *F*(*KeyWord*, *Document*) is appearing frequency of keyword *KeyWord* in document *Document*, *N* is total amount of training texts, n is the number of documents that contain keyword *KeyWord* in training samples, *Weight*(*KeyWord*, *Class*) is weight of category about class *Class* for keyword *KeyWord*. This paper gets standard document vector that stands for topics through formula above, every value of vector is the corresponding value of weight for keywords, and the number of dimension is the number of keywords for vector.

4.2 The extraction layer of Web news elements

The extraction layer of Web news elements is mainly responsible for obtaining the text structurally, which includes the Web news URL, title, time of releasing, source of releasing, content, hyperlinks and other text according to initial *NewsSet*, initial *UrlSet* and *WaitingUrlQueue*. The results extracted are organized into a Web news corpus, which is used in the filter layer of Web news elements.

In order to improve the extraction precision and efficiency of Chinese Web news in the design of this layer, this model uses open source library NekoHtml that parses HTML webpages [12], converts the data of webpages to plain text format, locates Web news title in <*title*> label pertinently through analysing the organizational structure characteristics of Web news, locates on time of releasing in the next line of Web news title, and extracts source of releasing in adjoining element of Web news releasing time.

Through analysing the structure of the Web news HTML label including user navigators, floating ads, special theme menus, friendly link embedded in the webpages, it can be inferred that Web news content is made up of numerous natural paragraphs, each natural paragraph contains several Chinese punctuations, so this model uses the corresponding regular expression to eliminate the disturbance of noise object and determine whether the extracted information is Web news content or not.

4.3 The filter layer of Web news elements

The filter layer of Web news elements is mainly responsible for calculating the relativity of Web news content and the relativity of Web news hyperlinks. The calculation results of Web news content relativity are reorganized into the Web news corpus, which is used in the application layer of Web news extraction results. The calculation results of Web news hyperlinks relativity are put into *InitialUrlQueue*, which is used in the extraction layer of Web news elements. In this layer, this paper elaborates mainly three algorithms in order to achieve this process of filtering Web news elements.

In order to insure high relativity of webpages extracted and high relativity of hyperlinks reserved related to topics, this paper analyses them utilizing the method of filtering low relevant or irrelevant webpages and hyperlinks related to topics. The method proposed in this paper completes relativity calculation from two aspects of webpage content and network topology structure referring to three algorithms.

4.3.1 The relativity algorithm based on analysing webpage content

The relativity algorithm based on analysing webpage content is mainly responsible for calculating webpage relativity using the characteristics of webpage content, the general method is vector space model [13]. Traditional vector space model can complete relativity computation using initial standard topic vector and a given value of threshold [14]. Although this method is concise and clear, it ignores the feedback and guiding effect of subsequent extracting content related to topics, so this paper adds self-adaptation method on the basis of traditional vector space model, the relativity calculation formula of document *Document* and standard vector *Vector* is shown as follows.

As shown in Eq. (2), $Vector_{KeyWord}$ is a value of topic standard vector *Vector*, this method can adjust the standard vector and value of threshold automatically according to the information of follow-up feedback in self-adaptation stage. This adjustment is not conducting every time as completing analysis of a webpage, but has a certain interval.

Improving the value of threshold can get higher precision about content extracted, but reducing the value of threshold can get wider extracting range in the situation of lower webpage topic relativity. Sum(*T*), which expected in the *T* interval, is the number of documents extracted, Sum(*T*1) is the number of documents extracted in *T* time,

Sum($T2$) is total number of document is extracted in T time, Sum($T3$) is the number of documents extracted in T time related to topics, Sum($T4$) is total number of documents extracted in T time related to topics, the strategy of threshold value adjustment proposed in this paper is shown in Algorithm1. The process of modifying standard vector needs pass through continuous analysis for Web news extracted, in order to extract the new characteristic vector.

$$Relativity(Document, Vector) = \frac{\sum_{KeyWord} Weight(KeyWord, Document) * Vector_{KeyWord}}{\sqrt{\left(\sum_{KeyWord} Weight(KeyWord, Document)^2\right)\left(\sum_{KeyWord} Vector_{KeyWord}^2\right)}}. \quad (2)$$

Algorithm1: Calculate_WebpageContent_Relativity

Input: $WebNewsDB=\{r_1, r_2, \ldots, r_{i-1}, r_i, r_{i+1}, \ldots, r_n\}$, $Vector_{standard}$, Threshold, α, μ, β, γ, *InitialTime*, T, Sum(T).

Output: URListRT, URListIRT.

Calculate_WebpageContent_Relativity(WebNewsDB, r[], SystemData s)

//The array object of WebNewsDB stores elements of Web news, the object of SystemData stores member variables.

```
Begin
    For i=0 to r.size()-1 Do
        relativity=Relativity(r[i].content,s.getVector(
        ));
        //Calculate the relativity between the content
        of r[i] and standard vector using Relativity
        method.
        If getCurrentTime()-
        s.getInitialTime()>=s.getT() Then
            //Adjust the value of threshold according
            to strategy.
            If getSum(r,s.getT())>s.getSum() and
            getTotalSum(r,s.getT())>s.getSum()*s.getT()
            Then
                s.setThreshold(Threshold* α);
            ElseifgetTopicSum(r,s.getT())<getSum(r,
            s.getT())*μ and
            getSum(r,s.getT())>s.getSum() and
            getTotalSum(r,s.getT())>s.getSum()*s.getT()
            Then
                s.setThreshold(Threshold* β);
            ElseifgetSum(r,s.getT())<s.getSum() and
            getTotalSum(r,s.getT())<s.getSum()*s.getT()
            Then
                s.setThreshold(Threshold* γ);
            End If
            s.setInitialTime(getCurrentTime());
            //Reset InitialTime.
        End If
        If relativity>=s.getThreshold() Then
            URListRT.addlist(r[i].url);//The Web
            news related to topics
        Else
            URListIRT.addlist(r[i].url);//The Web
            news unrelated to topics
        End If
    End For
    calVector(URListRT);
    //Calculate standard vector viewed as keyword
 vector using calVector method.
    adjDB(URListRT);
    //Web news content related to topics is reorganized
 into the Web news corpus.
    End
```

4.3.2 The relativity algorithm based on webpage hyperlinks

When the topic-focused web crawler extracts webpages, it also extracts linking-out URLs of webpage to ensure its recycling execution except for processing the webpage content. However, the relativity of these linking-out URLs differs greatly in certain situations, so this paper analyses hyperlinks extracted to improve the extracting efficiency using *PageRank* algorithm.

The *PageRank* algorithm, which is a query-independent algorithm [15], is a frequently-used method to calculate the importance of linking-out URLs, its thinking is that more linked-in URLs number of one webpage, more important it is [16]. At the same time, the quality of linked-in URLs also has an effect on its importance. The calculation formula of authority *Priority*(I) for webpage I is shown as follows.

As shown in Eq. (3), D is an attenuation factor between zero and one, $1-D$ denotes the authority degree of each webpage, so D parts of each webpage authority degree are passed actually, $B(I)$ is a set of all webpages that point to webpage I, *Number*(I) denotes the number of linking-out URLs in webpage I.

Whenever the crawler extracts a new webpage, it will extract the URLs contained, so the corresponding priority of *InitialUrlQueue* also should be calculated again, in order to decrease the workload of calculation, the webpages downloaded are buffered in the process of extracting, the priority of *InitialUrlQueue* will be recalculated when only a certain interval reaches, the relativity algorithm based on webpage hyperlinks proposed in this paper is shown as follows taking Web news as processing object.

Algorithm2: Calculate_WebpageHyperLink_Relativity

Input: $WebNewsDB=\{r_1, r_2, \ldots, r_{i-1}, r_i, r_{i+1}, \ldots, r_n\}$, *InitialTime*, T.

Output: InitialUrlQueue.

Calculate_WebpageHyperLink_Relativity(WebNews DB, r[], SystemData s)

//The array object of WebNewsDB stores elements of Web news, the object of SystemData stores member variables.

```
Begin
    For i=0 to r.size()-1 Do
        HyperLinkList=convert(r[i].hyperlinks);
```

```
        //Store hyperlink instances of r[i] to object of list.
        relativity=Priority(HyperLinkList.get(j));
        //Calculate the relativity of hyperlink instance using Priority method.
        HyperLinkList.get(j).set(relativity);
        Extracting_Strategy(HyperLinkList.get(j));
        //Extracting_Strategy method is shown in Algorithm3.
        End For
            For j=0 to HyperLinkList.size()-1 Do
        End For
        If getCurrentTime()-s.getInitialTime()>=s.getT() Then
            ReOrganize(InitialUrlQueue);
            //Adjust the relativity of every hyperlink instance in InitialUrlQueue.
            s.setInitialTime(getCurrentTime());
            //Reset InitialTime.
        End If
    End
```

$$Priority(I) = (1-D) + D \times \sum_{J \in B(I)} \frac{Priority(J)}{Number(I)}. \quad (3)$$

4.3.3 The design of extracting strategy

The topic-focused information appointed by users usually takes a very small part in whole network information, so the expected requirements cannot be met either in efficiency or in recall as searching information according to the traditional breadth-first or depth-first method [17]. When the topic-focused web crawler extracts information along a specific direction, it usually barges up against a channel blocked, which means that the content of current webpage is irrelevant to the topic or its relativity is less than the value of threshold, although it will find other channel instead of current one, it can cause the situation that webpages will be discarded together in deeper layers of blocked channel [18], in most circumstances, some of these webpages are also related to topics. Based on analysis above, this paper proposes a webpage extracting strategy on the basis of gene factor, which is shown as follows taking Web news as processing object.

This strategy sets same default value Val of topic relativity viewed as priority for hyperlinks, because the hyperlinks of *InitialUrlQueue* have conducted strict filter, they have a high relativity with the topic, the value Val set is greater than the subsequent value Val got after relativity calculating. On the other hand, larger priority, which is set for *InitialUrlQueue*, can still be updated having precedence over subsequent webpages.

Algorithm3: Extracting_Strategy

Input: $WebNewsDB=\{r_1, r_2, \ldots, r_{i-1}, r_i, r_{i+1}, \ldots, r_n\}$, ρ, σ, *InitialTime*, *T*.

Output: UrlQueue, ExtQueue.

```
Extracting_Strategy(HyperLinkList hyperlink)
//The object of HyperLinkList stores a hyperlink instance.
Begin
    SystemData s;//The object of SystemData stores member variables.
    parentval=findParent(hyperlink.getLink);
//Find the parent webpage's value Val of current hyperlink.
    If hyperlink.getPriority()>=ρ Then
        hyperlink.setPriority(parentval);
    Else
        hyperlink.setPriority(parentval* σ);
    End If
    InitialUrlQueue.enqueue(hyperlink);
    If getCurrentTime()-s.getInitialTime()>=s.getT() Then
        SortByVal(InitialUrlQueue);
        //Sort data elements in InitialUrlQueue according to value Val.
        s.setInitialTime(getCurrentTime());
        //Reset InitialTime.
    End If
    WaitingUrlQueue.enqueue(InitialUrlQueue.dequeue());
    //Put the hyperlink positioned in the front of InitialUrlQueue into WaitingUrlQueue.
    Perform_Extraction(WaitingUrlQueue);
    //Performs the extracting process of Web News instances.
End
```

This strategy can ensure the execution of extracting on the appointed main channel all the time using initial relativity value Val and relativity value Val got by webpage hyperlink analysis calculation, when main channel is blocked, this strategy can create a subchannel from the main channel, in which the process of extracting keeps going on, thus it avoids the problem of ignoring many other related webpages in order to get local optimization.

4.4 The application layer of Web news extraction results

The application layer of Web news extraction results is mainly responsible for mining potential value in the background of Web news elements extracted related to topics. Based on Web news extraction results got making use of the model, algorithms and technology in this paper, researchers can further develop application oriented to requirement of users.

The researchers can develop application oriented to trust analysis of Web news based on Web news extraction results, although topic-focused Web news have been extracted exactly, however, in which some information communicated by web media is illusive [19], the application oriented to trust analysis of Web news should show the degree of trust. The researchers can develop application oriented to currency analysis of Web news

based on Web news extraction results, although topic-focused Web news have been extracted exactly, however, in which some information communicated by web media is outdated [20], the application oriented to currency analysis of Web news should show the degree of currency. The researchers can develop application oriented to hot topic detection of Web news based on Web news extraction results, although topic-focused Web news have been extracted exactly, some patulous information communicated by web media, which can reorganize a new hot topic, is concealing in Web news [21], the application oriented to hot topic detection of Web news should mine new hot topic from Web news extraction results related to appointed initial topics by users. The researchers can develop application oriented to topic evolution tracking of Web news based on Web news extraction results, although topic-focused Web news have been extracted exactly and new hot topic can be detected from appointed initial topics by users or interrelating with it through developing application, however, it is not achieved to track topic evolution path [22], the application oriented to topic evolution tracking of Web news should show time line of topic evolution tracked.

5 The experimental results and analysis of model

This paper carries out experiments and analyses experimental results in order to validate feasibility, validity and superiority of the model proposed. In the process of completing experiments, the author adopts the experimental environment shown as follows. The processor is dual core, the memory is 32G, the language of computer programming design is Java, the platform of experimental design and implementation is MyEclipse, the platform of experimental data storage and management is SQL Server.

The author designs the experimental form of model based on Web news extraction model's design and description of function in each layer. This form uses the Matisse Form Class of MyEclipse platform as the top container including several modules. The first module has the function of importing a set of *NewsSet* from text file, excel file or database, which is used in the extraction layer of Web news elements. The second module has the function of importing a set of *UrlSet* from text file, excel file or database, which is used in the definition layer of Web news topics and the extraction layer of Web news elements. The third module has the function of defining a set of *TopKeyWordSet* and its corresponding weight value used in the definition layer of Web news topics, which denote keywords and its corresponding important degree of Web news topics extracted. The fourth module has the function of setting the value of parameters using introductory manner, which are mainly used in three important algorithms of calculating webpage content relativity, calculating webpage hyperlink relativity and extracting strategy. The fifth module has the function of showing Web news extraction results based on improved topic-focused web crawler key technology. The sixth module has the function of prospecting application development based on Web news extraction results, such as the application oriented to trust analysis of Web news, currency analysis of Web news, hot topic detection of Web news, hot topic evolution tracking of Web news and so on.

5.1 The experimental results of model

Based on the model and realization process of core algorithm presented in this paper, the author conducts a detailed expatiation taking MH17 airliner event and its related Web news as the case of application.

In the form of importing a set of *NewsSet* from text file, excel file or database, the users can respectively click three buttons i.e. From Text, From Excel and From DataBase in order to import website URLs containing Web news stored in text file, excel file or database. The users can also input appointed website URLs containing Web news in jTable component. In the end, the users should click button that is To *NewsSet* in order to store these URLs used in the extraction layer of Web news elements into *NewsSet*, which is shown in Fig. 2.

In the form of importing a set of *UrlSet* from text file, excel file or database, the users can respectively click three buttons i.e. From Text, From Excel and From DataBase in order to import instance URLs related to Web news topics stored in text file, excel file or database. The users can also input appointed instance URLs related to Web news topics in jTable component. In the end, the users should click button that is To *UrlSet* in order to store these URLs used in the definition layer of Web news topics and the extraction layer of Web news elements, which is shown in Fig. 3.

In the form of defining a set of *TopKeyWordSet* and its corresponding value of weight, the users can input several keywords related to Web news topics and assign its corresponding instructional value of weight in jTable component. In the end, the users should click button that is To *TopKeyWordSet* in order to store these data used in the definition layer of Web news topics, which is shown in Fig. 4.

In the form of setting the value of parameters, the users can respectively select the value of parameters in the background of experimental guidance through jComboBox and jListBox component, which include α, μ, β, γ, Threshold, Sum(T) applied in the calculating webpage content relativity algorithm, include ρ, σ, Val applied in the extracting strategy algorithm, include *InitialTime*, T applied in the calculating webpage content relativity algorithm, calculating webpage hyperlink relativity algorithm and extracting strategy algorithm, include k used to control percentage of showing extraction results, the users should click button that is Set Parameter in order to store these parameters are used in the filter layer of Web news elements, which is shown in Fig. 5.

In the form of showing Web news extraction results, the users can click button that is EXTRACTING. The form of showing Web news extraction results can efficiently and accurately display the result of Web news related to topics combining three algorithms mainly, which is shown in Fig. 6.

In the form of prospecting application development, the users can respectively click four panels switched. When the users click panel of trust analysis, it will show the trust degree of Web news related to topics. When the users click panel of currency analysis, it will show the

currency degree of Web news related to topics. When the users click panel of hot topic detection, it will show new hot topic from Web news extraction results related to appointed initial topics by users. When the users click panel of hot topic evolution tracking, it will show time line of hot topic evolution tracked, which is shown in Fig. 7 taking panel of currency analysis as an example.

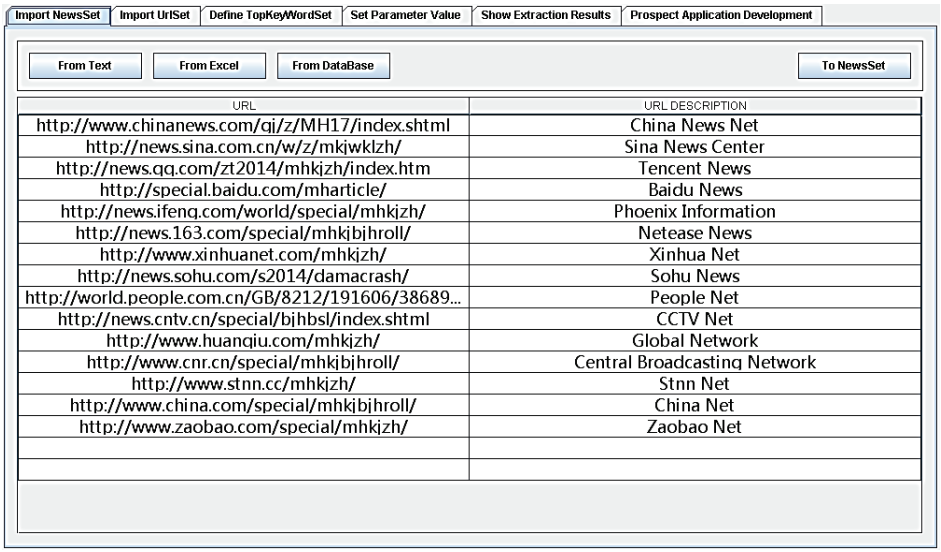

Figure 2 The form of importing a set of *NewsSet*

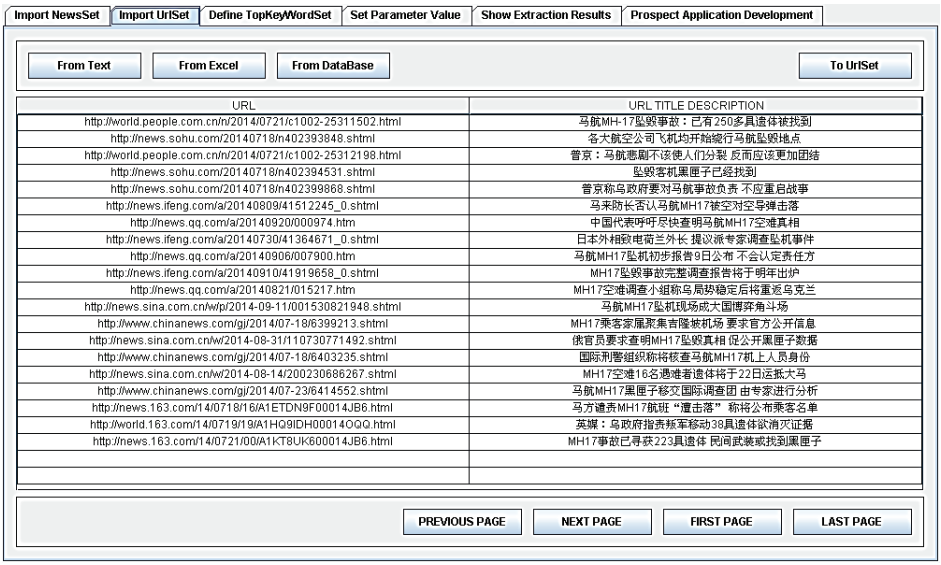

Figure 3 The form of importing a set of *UrlSet*

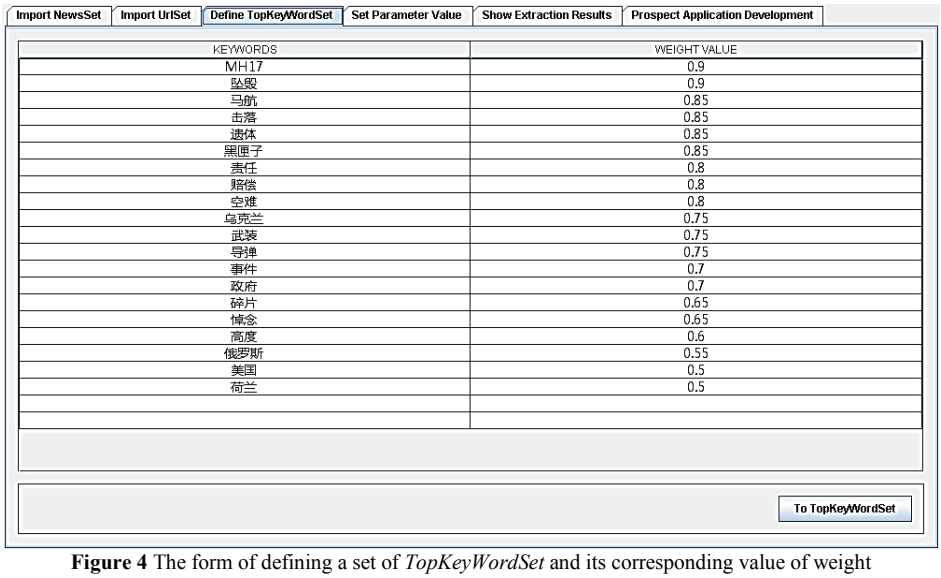

Figure 4 The form of defining a set of *TopKeyWordSet* and its corresponding value of weight

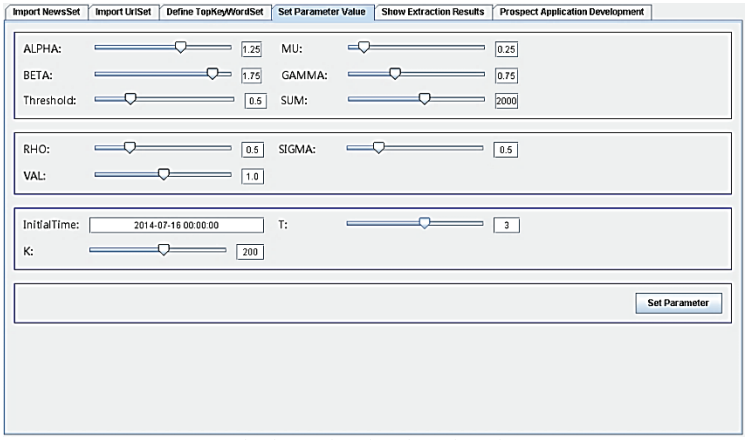

Figure 5 The form of setting the value of parameters

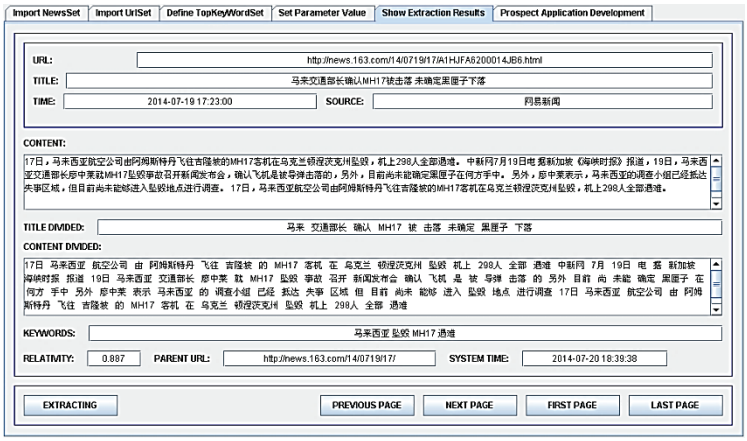

Figure 6 The form of showing Web news extraction results

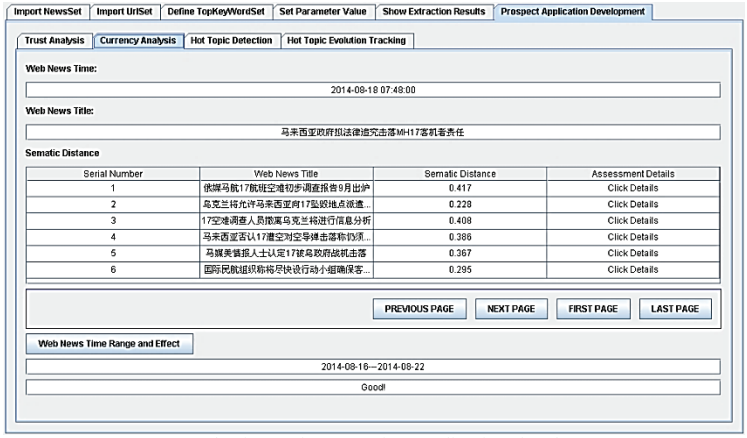

Figure 7 The form of prospecting application development

5.2 The experimental analysis of model

Based on the experimental process above, the author conducts a detailed analysis and discussion about accuracy, precision and flexibility of the model proposed in this paper. Table 1 shows the extraction results compared with the traditional method, which is a universal web crawler. It can be analysed that the universal web crawler has a wide extraction range, but the executive time is approximate. The main innovation of the model presented in this paper is that it analyses and calculates the relativity for Web news content and hyperlinks, filters some web pages, which are less than relativity value of threshold, so the gap of executive time will become smaller between the improved topic-focused web crawler and the universal web crawler with the growth of searching depth.

The algorithms of the model proposed in this paper are compared with the best first search algorithm taking MH17 airliner event and its related Web news as the extraction object of Web news topics from precision and recall. The result of experimental comparison is shown in Fig. 8 and Fig. 9.

As shown in Fig. 8, the experimental precision of algorithms proposed in this paper is close to the parallel comparing with best first search algorithm with growth of Web news quantity, whose precision is a little high, but as shown in Fig. 9, the experimental recall of algorithms proposed in this paper has its outstanding superiority. In the situation of extracting few Web news related to topics,

due to high relativity of the main channel opened up through defining the initial Web news URLs, the recall of algorithms is almost the same compared with best first search algorithm, but this efficiency of algorithms proposed in this paper is obviously higher in latter process of extracting more Web news related to topics. The reason of existing of this phenomenon is that the usage of improved method can locate the relevant webpages accurately; on the other hand, improved topic-focused web crawler algorithm can find a lot of webpages abandoned, in this process, further reflect accuracy, precision and flexibility of the model directly. The advantages of algorithms proposed in this paper have also been materialized in extraction results comparison.

Table 1 The comparison of extraction results between the improved topic-focused web crawler and the universal web crawler

The category of web crawler	Webnews searched	Webnews related to topics	Webnews unrelated to topics	Webnews unsearched	The searching time (s)	The extracting time (s)
The universal web crawler	4685	2954	1731	140	1044	1328
The improved topic-focused web crawler	3373	3357	16	84	773	830

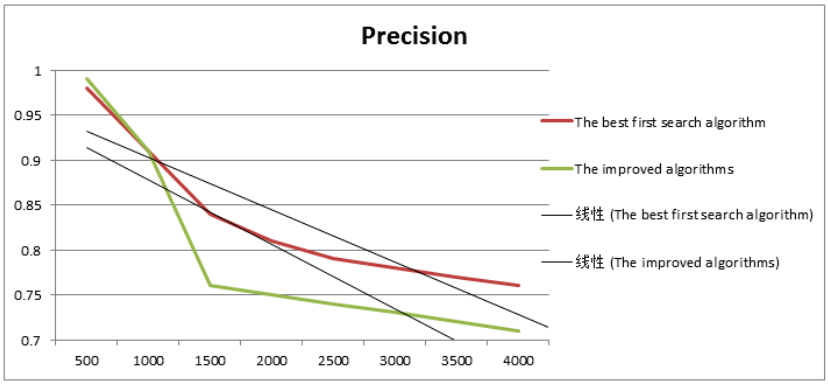

Figure 8 The comparison of experimental precision

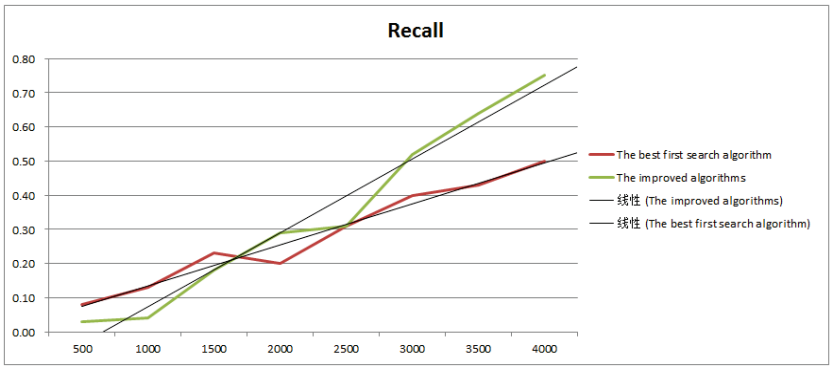

Figure 9 The comparison of experimental recall

6 Conclusion

This paper completes a research on model of network information extraction based on improved topic-focused web crawler key technology, which takes topic-focused web crawler key technology as a research core and executes the process of topic definition, data acquirement, data analysis, data filtering, data storage and data application taking Web news as research object from the point of innovation. In the process of model research and implement, this paper proposes three important algorithms of calculating webpage content relativity, calculating webpage hyperlink relativity and extracting strategy in order to eliminate shortcomings existing in traditional method. The experiment and its analysis results of model do key contributions for the feasibility, validity and superiority of network information extraction request, improve the efficiency of coordinating network information for users, enhance the availability of websites, build scientifically and improve service functions of websites, and improve business operational efficiency and clicking rate of website. In a word, the process of design, research and implement has a certain practical application value, which establishes the real and exact foundation of dataset for continual research and application on Web data mining direction.

7 Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 71271209, the project of science and technology plan of Beijing Education Committee under Grant Nos. KM201311417011, Funding Project for Academic Human Resources Development in Beijing Union University under Grant Nos. BPHR2012A02, the project of Philosophy and Social Science Planning of Beijing Nos. 13JGC090, and the project of a new starting point in Beijing Union University.

8 References

[1] Chen Mo; Yang Xiaoping; Liu Ting. A research on user behavior sequence analysis based on social networking service use-case model. // International Journal of u- and e-Service, Science and Technology. 7, 2(2014), pp. 1-4. DOI: 10.14257/ijunesst.2014.7.2.01

[2] Zhang Ji; Li Hongzhou; Gao Qigang; Wang Hai; Luo Yonglong. Detecting anomalies from big network traffic data using an adaptive detection approach. // Information Sciences. 6, 3(2014), pp. 96-97.

[3] Pandey Suraj; Nepal Surya. Cloud Computing and Scientific Applications-Big Data, Scalable Analytics, and Beyond. // Future Generation Computer Systems. 29, 7(2013), pp. 1774-1775. DOI: 10.1016/j.future.2013.04.026

[4] Zhu Zhiguo. A novel method for discovering frequent changing patterns from historical web access data. // ICIC Express Letters, 8, 9(2014), pp. 2443-2445.

[5] Wu Jiagao; Zhou Fankun; Zhang Xueying. Research of the Extraction Method of Event Properties Based on the Combining of HMM and Syntactic Analysis. // Journal of Nanjing Normal University (Natural Science Edition). 32, 1(2014), pp. 30-34.

[6] Zhang Hongli; Huang Shouming. Web Information Extraction Method Based on MapReduce. // Journal of Anhui Science and Technology University. 27, 2(2013), pp. 72-74.

[7] Li Wen; Zheng Bangxi; Deng Wu. Research on Web Information Extraction Model Based on XML and DOM Technologies. // Journal of Dalian Jiaotong University. 34, 3(2013), pp. 96-98.

[8] Du Yajun. Research on Collaborating Strategy among the Multi-agent Focused Crawlers. // Journal of Xihua University (Natural Science), 32, 1(2013), pp. 31-33.

[9] XieZhijun; Yang Wu; Li Zhiying; Song Jingjing. Research on Focused Crawler Based on HMM. // Journal Chongqing Technology Business University (Natural Science Edition). 29, 2(2012), pp. 66-68.

[10] Bai Yuzhao; Liang Jiuzhen. Research and Implementation for Focused Crawler Based on Probabilistic Mode. // Computer Engineering and Science. 35, 1(2013), pp. 160-162.

[11] Nasomyont, Tamrerk. A study on the relationship between search engine optimization factors and rank on google search result page. // Advanced Materials Research. 3, 4(2014), pp. 1462-1464. DOI: 10.4028/www.scientific.net/AMR.931-932.1462

[12] Chen Xuegang. Research and realization of E-commerce monitor system based on focused web crawler. // Information Technology Journal. 12, 17(2013), pp. 4033-4035. DOI: 10.3923/itj.2013.4033.4039

[13] Min Jun-Ki. Combining localized fusion and dynamic selection for high-performance SVM. // Expert Systems with Applications.42, 1(2014), pp. 9-11.

[14] Chen Mo; Yang Xiaoping; Sun Meng; Zhao Yun. Research on model of network information currency evaluation based on web semantic extraction method. // International Journal of Future Generation Communication and Networking. 7, 2(2014), pp. 103-105. DOI: 10.14257/ijfgcn.2014.7.2.11

[15] Guo Yi; Chen Hao. Microblog user ranking based on PageRank and Hadoop. // WIT Transactions on Information and Communication Technologies. 49, 1(2014), pp. 1083-1085.

[16] Sakakura, Yuta. An improved method for efficient PageRank estimation. // Lecture Notes in Computer Science, 8645, 2(2014), pp. 208-210. DOI: 10.1007/978-3-319-10085-2_19

[17] Balla, Andoena. Real-time web crawler detection. // 18th International Conference on Telecommunications, 2011, pp. 428-430. DOI: 10.1109/cts.2011.5898963

[18] Ahmadi-Abkenari, F. A clickstream-based web page significance ranking metric for web crawlers. // 5th Malaysian Conference in Software Engineering, 2011, pp. 223-225. DOI: 10.1109/mysec.2011.6140674

[19] Zhan Zhijian; Lin Feng; Yang Xiaoping. Semantic Similarity Calculation of Short Texts Based on Language Network and Word Semantic Information. // Communications in Computer and Information Science. 451, 1(2014), pp. 215-217. DOI: 10.1007/978-3-662-44491-7_17

[20] Zhang Yaming; Tang Chaosheng. Information propagation model based on the dynamics of complex networks in mircoblogging. // Journal of Computational Information Systems. 10, 1(2014), pp. 443-445.

[21] Zhu Tao; Lin Yumin; Cheng Ji; Wang Xiaoling. Efficient diverse rank of hot-topics-discussion on social network. // Lecture Notes in Computer Science. 8485, 1(2014), pp. 522-524.

[22] Lu Ran;XueSuzhi; Ren Yuanyuan; Zhu Zhenfang. A modified approach of hot topics found on micro-blog. // Lecture Notes in Electrical Engineering. 269, 1(2013), pp. 603-605.

Authors' addresses

Mo Chen
School of Information, Renmin University of China,
Beijing 100872, China
Business College of Beijing Union University,
Beijing 100025, China
mo.chen@buu.edu.cn
chenmoky@sohu.com

Xiao-Ping Yang
School of Information, Renmin University of China,
Beijing 100872, China
chenmokb@ruc.edu.cn